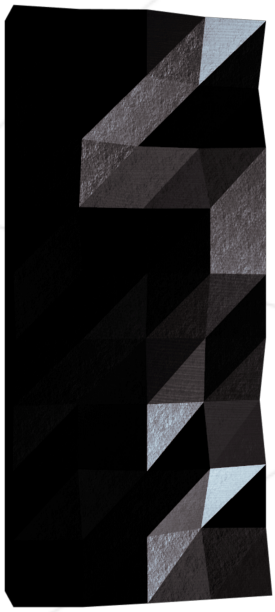


Training Piscine datascience - 3

The present

Summary: Today, you will see understanding the present

Version: 1.00

Contents

Chapter I

General rules

- You have to render your modules from a computer in the cluster either using a virtual machine:
    - You can choose the operating system to use for your virtual machine
    - Your virtual machine must have all the necessary software to realize your project. This software must be configured and installed.
- Or you can use the computer directly in case the tools are available.
    - Make sure you have the space on your session to install what you need for all the modules (use the goinfre if your campus has it)
    - You must have everything installed before the evaluations
- Your functions should not quit unexpectedly (segmentation fault, bus error, double free, etc) apart from undefined behaviors. If this happens, your project will be considered non functional and will receive a `0` during the evaluation.
- We encourage you to create test programs for your project even though this work **won't have to be submitted and won't be graded**. It will give you a chance to easily test your work and your peers' work. You will find those tests especially useful during your defence. Indeed, during defence, you are free to use your tests and/or the tests of the peer you are evaluating.
- Submit your work to your assigned git repository. Only the work in the git repository will be graded. If Deepthought is assigned to grade your work, it will be done after your peer-evaluations. If an error happens in any section of your work during Deepthought's grading, the evaluation will stop.
- By Odin, by Thor ! Use your brain !!!

Chapter II

Introductions

Data Scientists often use techno such as python, Jupyter Notebook, Julia ...

It is up to you to find the tools that suit you. You are free to use any language of your choice for this module.

The role of the data scientist is to predict "the future" with automatic learning models on past data, he must be a force of proposal to explain the possible interest to the implementation of his models, create tools to help decision making.

You are a fan of star wars, and you want to know if we should have predicted before if Anakin was going to burn on the dark side, you got the data of all the knights so go analyze them.
To do this, you have a data set containing all the knight's skills (the features) and a "knight" column (the target) that tells you which side of the force the knight is on.

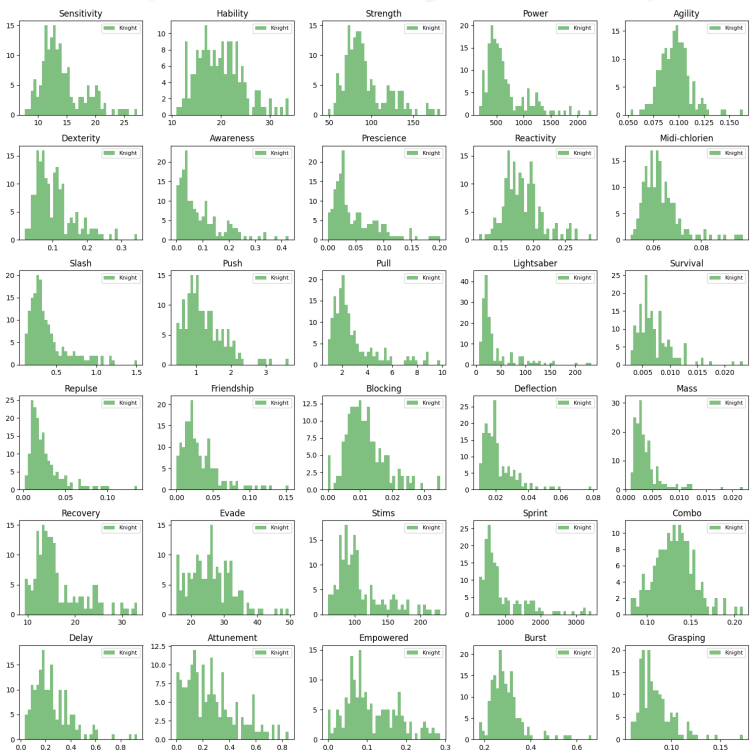
Chapter III

Exercise 00

	Exercise 00
Exercice 00 : Histogram	
Turn-in directory : $ex00/$	
Files to turn in : `Histogram.*`	
Allowed functions : `All`	

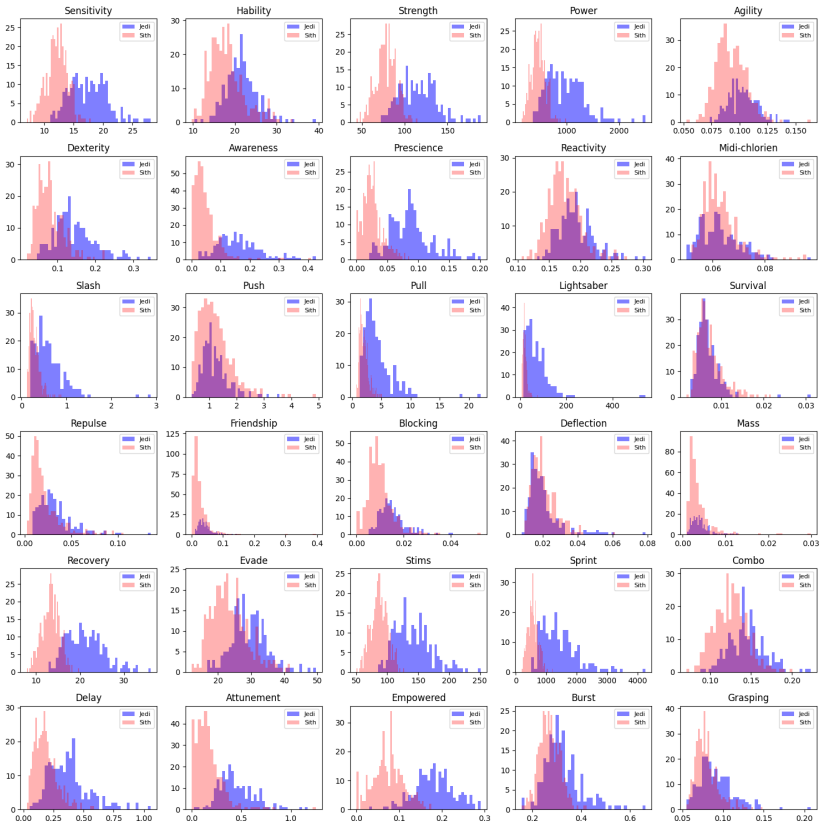
- Create a graph to visualize the data in the same histogram with the "Test_knight.csv" file.

For exemple:



Create a graph to understand the interaction between knight's skills (the features) and a "knight" column (the target) in the same histogram with the file "Train_knight.csv"

For exemple:



Chapter IV

Exercise 01

	Exercise 01
Exercice 01 : Correlation	
Turn-in directory : $ex01/$	
Files to turn in : `Correlation.*`	
Allowed functions : `All`	

- Write a Correlation Factors, to understand the collones with the most chorelation between the target (column "knight") and the features (all the other columns).

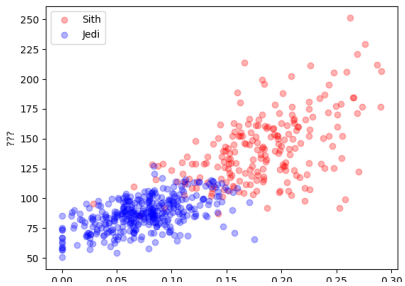
```
knight          1.000000
Empowered       0.793566
Stims           0.782914
Prescience      0.776614
Recovery        0.776454
Strength        0.742636
Sprint          0.733825
Sensitivity     0.730029
Power           0.708984
Awareness       0.696360
Attunement      0.659610
Dexterity       0.596534
Delay           0.590998
Slash           0.567134
Force           0.556141
Lightsaber      0.548236
Evade           0.456903
Combo           0.421465
Burst           0.416294
Hability        0.415185
Blocking        0.408042
Agility         0.358560
Reactivity      0.330499
Grasping        0.323872
Repulse         0.292999
Friendship      0.253730
Mass            0.077972
Survival        0.067016
Midi-chlorien   0.012838
Push            0.008303
Deflection      0.006522
```

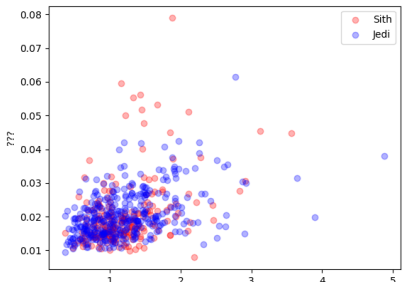
Chapter V

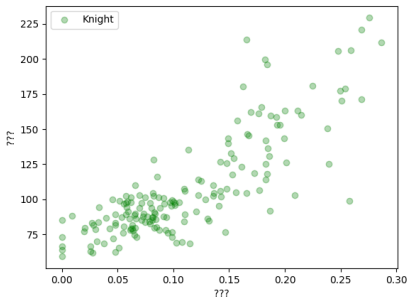
Exercise 02

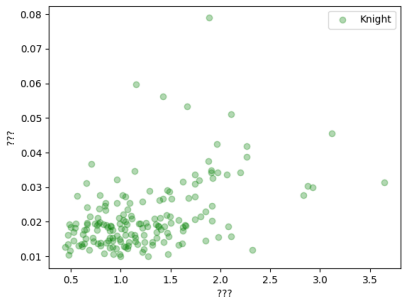
	Exercise 02
Exercice 02 : it's raining cats no points!	
Turn-in directory : $ex02/$	
Files to turn in : `points.*`	
Allowed functions : `All`	

- You have to display 4 graphs like the ones below (with Train_knight.csv and Test_knight.csv)
- One of the two graphs must visually separate the clusters, whereas the second one should mix them for each file









Chapter VI

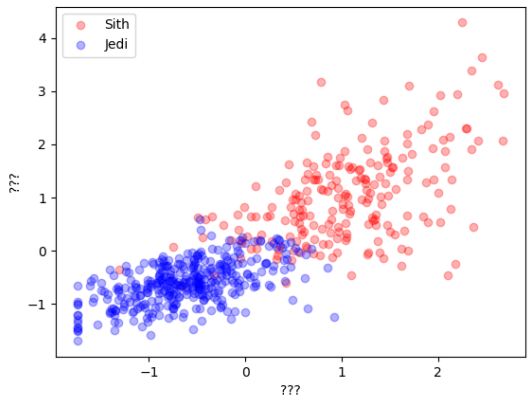
Exercise 03

	Exercise 03
Exercice 03 : standardization	
Turn-in directory : $ex03/$	
Files to turn in : `standardization.*`	
Allowed functions : `All`	

- standardize and print your data
- Display one of the graphs from the previous exercise with the standardized data.
- It must work with Train_knight.csv and Test_knight.csv

```
$>./standardization.*
Sensitivity Hability Strength ... Empowered Burst Grasping
     17.99    10.38   122.80 ...     0.26   0.46     0.11
     ...

Sensitivity Hability Strength ... Empowered Burst Grasping
      1.09    -2.07     1.26 ...     2.29   2.75     1.93
     ...
$>
```



Chapter VII

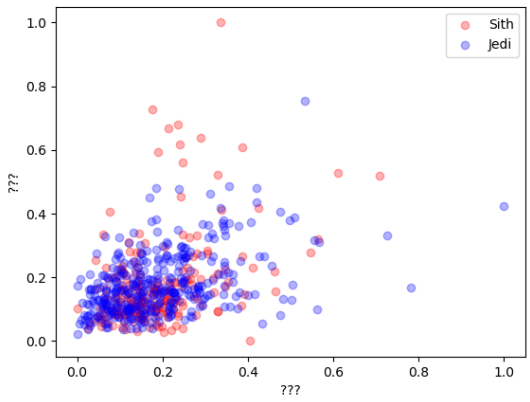
Exercise 04

	Exercise 04
Exercice 04 : Normalization	
Turn-in directory : $ex04/$	
Files to turn in : `Normalization.*`	
Allowed functions : `All`	

- Normalize and print your data
- Display the other graphs from exercise 02 with the normalized data
- It must work with Train_knight.csv and Test_knight.csv

```
$>./standardization.*
Sensitivity Hability Strength ... Empowered Burst Grasping
     17.99    10.38   122.80 ...     0.26   0.46     0.11
     ...

Sensitivity Hability Strength ... Empowered Burst Grasping
      0.52     0.02     0.54 ...     0.91   0.59     0.41
      ...
$>
```



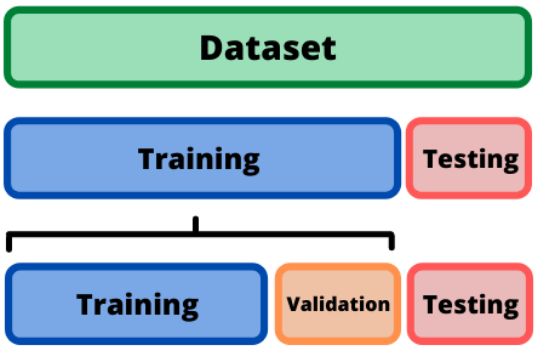
Chapter VIII

Exercise 05

	Exercise 05
Exercice 05 : Split	
Turn-in directory : $ex05/$	
Files to turn in : `split.*`	
Allowed functions : `All`	

- You have to write a program that randomly splits the file Train_knight.csv into Training_knight.csv and Validation_knight.csv
- You must be able to explain how many % you keep in each file and why

```
$>./split.* Train_knight.csv
$> ls
./split.* Train_knight.csv Training_knight.csv Validation_knight.csv
$>
```



Chapter IX

Submission and peer-evaluation

Turn in your assignment in your `Git` repository as usual. Only the work inside your repository will be evaluated during the defense. Don't hesitate to double check the names of your folders and files to ensure they are correct.

`The evaluation process will happen on the computer of the evaluated group.`